International population-based health surveys linked to outcome data

International population-based health surveys linked to outcome data: A new resource for public health and epidemiology

by Stacey Fisher, Carol Bennett, Deirdre Hennessy, Tony Robertson, Alastair Leyland, Monica Taljaard, Claudia Sanmartin, Prabhat Jha, John Frank, Jack V. Tu, Laura C. Rosella, JianLi Wang, Christopher Tait, and Douglas G. Manuel

Abstract

Background

National health surveys linked to vital statistics and health care information provide a growing source of individual-level population health data. Pooling linked surveys across jurisdictions would create comprehensive datasets that are larger than most existing cohort studies, and that have a unique international and population perspective. This paper's objectives are to examine the feasibility of pooling linked population health surveys from three countries, facilitate the examination of health behaviours, and present useful information to assist in the planning of international population health surveillance and research studies.

Methods

The design, methodologies and content of the Canadian Community Health Survey (2003 to 2008), the United States National Health Interview Survey (2000, 2005) and the Scottish Health Survey (SHeS) (2003, 2008 to 2010) were examined for comparability and consistency. The feasibility of creating common variables for measuring smoking, alcohol consumption, physical activity and diet was assessed. Sample size and estimated mortality events were collected.

Results

The surveys have comparable purposes, designs, sampling and administration methodologies, target populations, exclusions, and content. Similar health behaviour questions allow for comparable variables to be created across the surveys. However, the SHeS uses a more detailed risk factor evaluation for alcohol consumption and diet data. Therefore, comparisons of alcohol consumption and diet data between the SHeS and the other two surveys should be performed with caution. Pooling these linked surveys would create a dataset with over 350,000 participants, 28,424 deaths and over 2.4 million person-years of follow-up.

Conclusions

Pooling linked national population health surveys could improve population health research and surveillance. Innovative methodologies must be used to account for survey dissimilarities, and further discussion is needed on how to best access and analyze data across jurisdictions.

Keywords: population health, health surveillance, national health surveys

Authors: Stacey Fisher (stacey.fisher@uottawa.ca), Douglas G. Manuel (dmanuel@ohri.ca), Carol Bennett and Monica Taljaard are with the Ottawa Hospital Research Institute in Ottawa, Ontario. Stacey Fisher, Carol Bennett and Douglas G. Manuel are also with the ICES, as is Laura C. Rosella. Stacey Fisher, Monica Taljaard and Douglas G. Manuel are also with the School of Epidemiology and Public Health at the University of Ottawa, as is JianLi Wang. Douglas G. Manuel is also with Statistics Canada, as are Deirdre Hennessy and Claudia Sanmartin. Tony Robertson is with the Population and Public Health Research Group in the Faculty of Health Sciences and Sport at the University of Stirling in Scotland. Alastair Leyland is with the

MRC/CSO Social and Public Health Sciences Unit at the University of Glasgow. Prabhat Jha and Christopher Tait are with the Dalla Lana School of Public Health at the University of Toronto. Prabhat Jha is also with St. Michael's Hospital, as is Laura C. Rosella. John Frank is with the Scottish Collaboration for Public Health Research and Policy in Edinburgh, and the Usher Institute Centre for Population Health Sciences at the University of Edinburgh. JianLi Wang is also with The Royal's Institute of Mental Health Research, the Department of Psychiatry at the University of Ottawa, and the Mathison Centre for Mental Health Research and Education at the University of Calgary. Douglas G. Manuel is also with the Department of Family Medicine at the University of Ottawa and the Bruyère Research Institute in Ottawa. Jack V. Tu (deceased May 20, 2018) was with ICES, the Sunnybrook Schulich Heart Centre, and the Institute of Health Policy, Management and Evaluation at the University of Toronto.

What is already known on this subject?

- National population health surveys are key tools for understanding, monitoring and improving population health.
- Around the world, national population health surveys with similar objectives and designs are increasingly being linked at the individual level to vital statistics data and health care data, providing a valuable longitudinal perspective.
- Creating common variables may enable individual-level pooling of these linked surveys and could produce a new resource for population health research with an untapped international and population perspective.

What value does this study add?

- Health surveys in Canada, the United States and Scotland are largely comparable and common health behaviour variables can be constructed. However, Scotland uses a more detailed and standardized risk factor evaluation for alcohol consumption and diet data. Therefore, comparisons of alcohol consumption and diet data between Scotland and the other two countries should be performed with caution.
- Pooling national linked population health surveys is feasible and has the potential to be used for international health risk comparison, equity analysis, disease burden estimation and ongoing surveillance.
- Challenges introduced by survey dissimilarities will require innovative methodologies, and can be improved with the introduction of international standards for collecting core health-related measures. Jurisdictional data restrictions and privacy issues will also require discussion and resolution.

Introduction

In Canada and elsewhere, national population-based health surveys are increasingly being linked to vital statistics and health care data, bringing together large amounts of high-quality, nationally representative information about health risk factors with individual-level health outcomes.[1–5] Health surveys in Canada and the United States alone have collected detailed sociodemographic and health behaviour information from over 1 million respondents since 1997, and have been linked to over 6 million person-years of mortality follow-up.[1,6] Because national health surveys often have similar surveillance objectives and designs, pooling data from linked population health surveys could create a new resource for health surveillance and research, with an unparalleled international and population perspective.

National population health surveys vs. traditional epidemiology studies

National population health surveys collect a broad range of information about health status, health behaviours and sociodemographic characteristics from a representative sample of a country's community-dwelling population. These surveys are a cornerstone of population health surveillance (Table 1) and have a population perspective—they use a sampling approach that is designed to produce a population-representative sample (in terms of sociodemographic characteristics). This sample is used to estimate the prevalence of health conditions and risk factors within the population. It is also used to monitor population trends; inform policy development, implementation and evaluation; inform decisions about health resource allocation; and assess progress toward national health goals.

National population health surveys are typically conducted at regular intervals (often yearly) to provide up-to-date snapshots of the population's health. In contrast, traditional epidemiology studies—studies that use conventions taught in most introductory epidemiology courses to investigate a specific exposure–outcome relationship—typically have an etiological focus, and often use convenience sampling.

Population health surveys do not usually collect longitudinal information about survey respondents, unlike most traditional epidemiology studies, which typically involve actively determining outcomes, often with repeat exposure assessment during follow-up. Typically, population health surveys determine only baseline exposures, through self-response, and most surveys do not ascertain temporal outcomes because of their cross-sectional nature. However, linking health surveys to outcome data, such as vital statistics and health care data, introduces a longitudinal perspective that greatly increases the surveys' utility.

In addition to population health surveillance, national health surveys are used for population health research since they collect information that is not available in administrative health files (e.g., health behaviours). These data are used by researchers to study the relationships between social determinants and health outcomes, to evaluate disease and risk factor burden, and to study the role of risk factor modification in prevention. These data are also used to assess the performance of the health care system across sociodemographic and economic groups, and across groups with varying levels of illness. Data are also used to inform the development of health policy. National health surveys are key tools for understanding, monitoring and improving population health.

Individual-level pooling of national population health surveys

Meta-analyses have long been used to summarize collections of traditional epidemiology studies, offering increased statistical power and more precise effect estimates. Individual-level pooling of linked population health surveys may confer similar benefits to population health questions, and could produce a valuable new resource for modern population health planning, including the use of population-level multivariable risk algorithms[7–11] and microsimulation[12,13] to project disease burden and evaluate risk reduction strategies.

Meta-analysis of traditional epidemiology studies

Meta-analysis is a statistical procedure used to summarize the results from multiple independent clinical trials or observational studies investigating a specific exposure–outcome association. The key value of meta-analysis is that it involves aggregating data from all relevant studies, which produces a quantitative summary of a body of research with higher statistical power and more precise effect estimates than the individual studies alone. Meta-analysis can be used to reconcile inconsistent results from previous studies, and to investigate rare diseases and uncommon or weak risk factors that individual studies were unable to investigate.[14–16]

Meta-analyses also offer the opportunity to produce new insights through the exploration of statistical heterogeneity. Statistical heterogeneity is present in a meta-analysis when the effect estimate of interest differs across the studies by more than can be accounted for by sampling variation. This can be caused by differences in study design, statistical methodology or study quality—leading to methodological heterogeneity—or by differences in exposure or outcome

definitions, or population characteristics—leading to clinical heterogeneity.[17] In all meta-analyses, it is important to identify the presence or absence of heterogeneity because aggregating studies with inconsistent results can lead to inaccurate or misleading conclusions.[17-21] However, heterogeneity can also be "our greatest ally"[20] since investigating its causes can lead to significant scientific and clinical results.[17,21]

Individual patient data (IPD) meta-analyses involve pooling and reanalyzing raw data from eligible studies.[22] These meta-analyses are considered the gold standard of systematic reviews.[23] Pooling and reanalysis allow for the standardization of participant inclusion and exclusion criteria, variable definitions, confounder adjustment, and modelling. This leads to more accurate summary effect estimates,[24] and makes it easier to investigate the influence of participant-level characteristics on the effect estimate, and to identify subgroups where risk factor associations may vary. Despite the substantial advantages over meta-analyses without individual-level data, IPD meta-analyses are not frequently performed since they require substantial cooperation and organization, data sharing, and advanced statistical expertise.[24,25]

Application to national population health surveys

The aggregation of linked international health surveys creates a valuable resource for modern population health care planning and research. Pooling and analyzing individual-level data from national population health surveys using methods similar to IPD meta-analysis could produce more accurate effect estimates with less statistical uncertainty. Additionally, investigating survey-level heterogeneity and subgroups could produce new insights. Survey aggregation could

produce improved comparisons of disease risk, burden and trends internationally; facilitate equity analyses; and support health policy and priority setting.

This paper's objectives are to examine the feasibility of pooling linked population health surveys from three countries, facilitate the examination of health behaviours, and present useful information to assist in the planning of international population health surveillance and research studies. Detailed comparisons of the design, methodologies and content of national health surveys from Canada, the United States and Scotland were performed. Common variables were constructed, and sample size and estimate outcome counts are provided.

Methods

Survey designs for the Canadian Community Health Survey (CCHS) (cycles 2.1 [2003], 3.1 [2005] and 4.1 [2007], and CCHS 2008),[26] the United States National Health Interview Survey (NHIS) (2000 and 2005)[27,28] and the Scottish Health Survey (SHeS) (2003, 2008 to 2010)[29–32] were examined for comparability. This involved evaluating survey content, target populations and exclusions, sampling and administration methods, sample size and response rates, and linkage. Survey year, inclusion of health behaviour topics of interest (e.g., the NHIS collects detailed diet information every five years) and availability of mortality linkage were considered to select the relevant survey cycles.

Questions on smoking, alcohol, physical activity and diet were identified, and question construction, response categorization and structure were compared. Health behaviours were the focus since they are important health risk factors that are collected in virtually all health surveys,

and because they are conceptually complex and are observed using different approaches. Health behaviour concepts were assessed for comparability, and existing variables were used to create new common variables. Common variables were constructed to achieve the highest level of detail possible in all surveys, which were assessed and discussed by three reviewers.

Public use files were used to obtain sex-specific sample size estimates of survey respondents aged 20 and older from the three countries. CCHS estimates were obtained through collaboration with Statistics Canada. Public use NHIS data were downloaded from the Centers for Disease Control and Prevention (CDC) website (www.cdc.gov). Public use SHeS data were obtained from the United Kingdom Data Service website (www.ukdataservice.ac.uk).

CCHS mortality estimates were obtained through collaboration with Statistics Canada. NHIS mortality estimates were obtained from public use NHIS files linked to the National Death Index, which is also available for download from the CDC's website. Mortality estimates for the SHeS were obtained through collaboration with Scotland's Information Services Division. Mortality follow-up data for the CCHS and the NHIS went to December 31, 2011, while SHeS follow-up data went to December 31, 2014. Research ethics approval was obtained from the Ottawa Health Science Network Research Ethics Board.

Results

Survey comparability

The CCHS, NHIS and SHeS are government-funded, cross-sectional household surveys designed to support national health surveillance efforts in Canada, the United States and Scotland,

respectively.[2,6,26] The CCHS was administered biennially from 2001 to 2007, and has been administered annually since 2008. The NHIS has been administered annually since 1957. The SHeS was administered in 1995, 1998 and 2003, and annually since 2008. Results of the comparability analysis are summarized in Table 2.

Content

Core questionnaires collect information about sociodemographic characteristics, health status, health care services and health determinants. Information about additional health topics of interest are collected in rapid response modules (CCHS), survey supplements (NHIS) and a rotating biennial module (SHeS). The SHeS also collects anthropometric measurements and blood, saliva and urine samples from a subsample of survey respondents.

Target population and exclusions

The CCHS, NHIS and SHeS have comparable target populations that include the non-institutionalized national population and exclude active members of the military, those in prison and long-term care facilities, and those living in some remote areas (CCHS and SHeS) or outside the country (NHIS). The CCHS also excludes those living on reserves. The CCHS collects information only for those aged 12 and older, while both the NHIS and the SHeS collect information on all individuals, regardless of age.

Sampling methods

Although the countries' populations vary (Canada has 37.6 million residents, the United States has 329 million residents, and Scotland has 5.3 million residents), similar multistage area sampling methods, designed to produce annual national-level data, are used in the CCHS, NHIS

and SHeS. The CCHS also produces annual estimates at the levels of the provinces, territories and 110 health regions. The SHeS produces health-board-level data every four years. The sample size of the NHIS is not large enough to provide state-level data with acceptable precision, but data can be evaluated over multiple survey years to obtain estimates.

For the CCHS, sampling was done by allocating the annual sample size among the provinces and territories according to their population size and number of health regions, and then further allocating the sample among the health regions. The NHIS sampling frame for the 2000 and 2005 surveys used 358 primary sampling units—within which two further sampling units were used—and involved oversampling of both Blacks and Hispanics. For the SHeS, each year's sample was clustered, and the four-year sample was unclustered. In 2008, 25 strata of area deprivation were used in the SHeS to produce estimates at the health-board level, allowing for the oversampling of deprived areas. All the surveys used sample weights to account for selection probabilities and non-response bias.

Administration methods

All the surveys used computer-assisted personal interviews administered by trained interviewers. Approximately half of the CCHS interviews were administered using computer-assisted telephone interviews.

Sample size and response rates

The CCHS was administered to 130,000 respondents every two years when it began in 2001. Since 2007, it has been administered to 65,000 respondents annually. The total adult response

rate was 81% in 2003 and 76% in 2008. The NHIS has been administered to approximately 30,000 adult respondents annually since 1997. Response rates from a non-conditional sample of adults were 72% in 2000 and 69% in 2005.[27,28] The SHeS has a much smaller sample size than both the CCHS and the NHIS. Until 2011, the SHeS surveyed between approximately 7,000 adults per cycle. Since 2011, it has surveyed approximately 4,500 adult respondents annually. From 2003 to 2010, response rates for eligible adults were between 55% and 60%.[29,32]

Available linkages

The CCHS has been linked to vital statistics data up to December 31, 2011, and to hospital discharge abstracts, with plans for further data linkages.[1] Access to these data is restricted to Statistics Canada and the Statistics Canada research data centres. The NHIS has been linked to the National Death Index, with follow-up to December 31, 2011. Information on accessing public use data files, feasibility data files and restricted-access data is available from the CDC (www.cdc.gov). The SHeS has been linked to mortality and health administrative databases, including hospitalizations,[2] with mortality follow-up to December 31, 2014. Access to these data can be requested from the Public Benefit and Privacy Panel for Health and Social Care (www.informationgovernance.scot.nhs.uk).

Question construction, response categorization and structure

Common variables were created to measure smoking, alcohol consumption, physical activity and diet in the three surveys (Table 3). The common variables for smoking and physical activity are comparable between the CCHS, NHIS and the SHeS. The common variables for alcohol and fruit and vegetable consumption are comparable between the CCHS and the NHIS. However, the

SHeS collects and reports alcohol consumption and diet information using more detailed and standardized measures. Therefore, comparisons of alcohol consumption and diet data between the SHeS and the other two surveys should be performed with caution. In the SHeS, alcohol consumption is reported using units of alcohol. This is not directly comparable with the CCHS and NHIS, which use the more subjective "number of drinks." Similarly, the SHeS collects detailed fruit and vegetable consumption information, including amount consumed, while the CCHS and NHIS collect only frequency information. In-depth descriptions of how survey similarities and differences influenced common variable creation, and how these differences may affect their interpretation, can be found online at https://osf.io/4rczm/.

Mortality linkage sample size estimates

Approximately 87%, 94% and 83% of CCHS, NHIS and SHeS respondents, respectively, who agreed to data sharing and linkage were successfully linked to national mortality data (Table 4). Among those successfully linked, 19,227 deaths occurred among CCHS respondents during 1.8 million person-years of mortality follow-up. Among NHIS respondents, 6,341 deaths occurred during almost half a million person-years of follow-up. Among SHeS respondents, 2,856 deaths occurred during 160,000 person-years of follow-up.

Discussion

National population health surveys are the largest population-based cohorts with information on health status, health behaviours, sociodemographic characteristics, health care use and health-related quality-of-life measures. Given the surveys' broad objectives, these data are well suited

for many purposes, especially when linked to health outcome data. Pooling linked health survey data could produce a new population health research resource.

There are two main benefits of pooling linked population health surveys at the individual level across jurisdictions. First, combined data have a larger sample with greater statistical power, which produces more precise effect estimates. This enables additional subgroup analyses and more detailed examinations of mediation and interaction effects. Increased sample size also allows for improved examination of uncommon or weak risk factors, and uncommon outcomes, such as cancer and many chronic diseases. Second, these data could improve the generalizability of study findings. Relationships between survey exposures and linked outcomes that are consistent across jurisdictions are potentially more robust, compared with inconsistent relationships. The investigation of inconsistent relationships can also lead to new insights, similar to the investigation of heterogeneity in meta-analyses. For example, the effect of health behaviours on mortality risk may be associated with country-level differences in socioeconomic inequality or access to health care services.[33]

Larger sample size and improved generalizability lead to many research and surveillance opportunities. For example, most international analyses rely on aggregated results from different sources, so many studies have difficulties considering sociodemographic variables and addressing mediation, interaction and exposure–outcome lag time.[8,34] Because health surveys typically include sociodemographic questions regarding education, work history, income, ethnicity and immigrant status, these data are well suited for investigating health risks from an equity perspective.

Combined health survey data also enhance the ability to monitor the relationship between survey exposures and outcomes. For example, there are concerns that the relationship between smoking and health outcomes has changed over time, given changes in smoking patterns and the composition of smoking products. An international, longitudinal investigation of this relationship is possible with pooled, linked international population health surveys.

Furthermore, pooled, linked international population health surveys could be used to produce improved international comparisons of disease burden estimates. Disease burden reporting requires information about risk factor prevalence, outcome counts and relative risk estimates associated with the exposure of interest. Most current disease burden estimates, including those from the Global Burden of Disease study,[35] use the aggregated data approach first described by Levin (1978).[36] With this method, aggregate measures of risk factor prevalence and outcome counts are obtained from independent sources. Risk factor prevalence is obtained from population health surveys, outcome counts are obtained from vital statistics data sources, and relative risk estimates are obtained from independent epidemiology studies to describe the association between the risk factor and outcome. However, national health surveys that have been linked to outcome data can be used as single data sources for these studies,[8] and pooling these data from multiple countries would allow for a standardized analysis methodology.

Limitations and challenges

One of the greatest challenges of pooling linked population health surveys from different countries is the heterogeneity caused by survey question dissimilarities. It was difficult and labour-intensive to create common health behaviour variables using the surveys from Canada,

the United States and Scotland for this study, and the variables created were less detailed and were not entirely comparable across all surveys.

Over time, the ascertainment of behavioural risk factors has become more consistent across countries, and an increasing number of validation studies exist that indicate acceptable ascertainment bias.[40] However, there is a need for more consistency. For example, despite international recommendations for smoking ascertainment that are used in over 100 countries,[41] the lack of smoking history information in the NHIS prevented the calculation of pack years—a more detailed measure of smoking behaviour than the categorical measure of "smoking status" that was created in this study. Changes to the CCHS also prevented differentiation between former drinkers and non-drinkers in cycle 4.1. In this study, even if a concept was present, the time frame over which the exposure was ascertained often varied. Furthermore, some questions were collected optionally by geographic region, and there were differences in variable definitions and classification.

The comparison of exposures in multiple health surveys is challenging, and health surveys are constantly changing. To help with this, "cchsflow," an open-source library to support the harmonization of CCHS variables across survey cycles, was developed.[37] This approach to variable harmonization can be extended to other international population health surveys, and can be used to harmonize variables both across cycles within a single survey and between surveys from different countries. Survey metadata also support harmonization by improving survey cataloguing. Survey metadata are available in Data Documentation Initiative format, an international metadata standard developed for this purpose.[38,39]

Another challenge is decreasing response rates. If participants systematically differ from those who do not participate, the survey sample will be non-representative of the target population, and valid inference will be impeded. Non-respondents are repeatedly found to have unfavourable health behaviours and excess mortality compared with respondents.[42–45] Data linkage can be used to assess and, potentially, adjust for non-response bias. The extent of non-response bias in the SHeS was evaluated by Gorman et al. (2014)[46] by comparing rates of all-cause mortality and alcohol-related harm among survey respondents and the general population. Incidence rates were found to be lower among survey respondents, with survey-to-population rate ratios of 0.69 for alcohol-related harm and 0.89 for all-cause mortality. They concluded that heavy drinkers were less likely to respond to the SHeS than moderate or light drinkers. This type of comparison of respondents and non-respondents can inform weighting and imputation procedures to adjust for non-response bias.

Approaches to combining cycles of population health surveys from a single country have been developed,[47] but approaches to pooling surveys from different countries are more complex because of differences in each country's survey design. Modified meta-analytic methods and techniques used by internationally pooled epidemiology cohort studies, such as the European Prospective Investigation into Cancer and Nutrition,[48] may be used. However, new methodologies will need to be established. Additionally, differences in the underlying survey populations may also prevent the estimation of a pooled effect estimate for an exposure–outcome effect of interest. However, this will not be the case for all effects, and investigations into the sources of this heterogeneity could also produce important new insights.

Lastly, the largest practical limitation to pooling international linked population health surveys is data access. The accessibility of outcome-linked health survey data varies across countries. For example, access to the mortality-linked NHIS data is publicly available on the CDC website. In contrast, access to the equivalent information in Canada, Scotland and many other countries is restricted. That said, unlinked health survey data are often publicly available—including data from the CCHS, which has a Statistics Canada Open Licence. The NHIS has demonstrated that it is possible to assess how to include linked outcomes to existing public use surveys, while ensuring there is no increase in re-identification risk and ensuring adherence to existing data sharing principles.

The Strategy for Patient-Oriented Research (www.cdp.hdrn.ca) was developed to address data access and harmonization efforts across Canada. A similar model could be used to facilitate analogous tasks for multi-country studies, including the pooling of linked population health surveys. Within networks such as the International Population Data Linkage Network (www.ipdln.org), there has also been more discussion and interest in conducting studies using data from multiple countries. Improvements to cross-jurisdictional data sharing and privacy issues are necessary for the benefits of pooled health survey analyses to be fully realized. This is beyond the scope of this paper.

Conclusion

The use of pooled national population health surveys linked to health outcomes has enormous potential for international health risk evaluation and comparison, equity analysis, disease burden

estimation, and ongoing surveillance. Innovative methodologies will be required to mitigate challenges introduced by survey dissimilarities, and these methodologies can be improved with the introduction of international standards for collecting core health-related measures. Jurisdictional data restrictions and privacy issues require discussion and resolution.

List of abbreviations

CAPI: computer-assisted personal interview

CCHS: Canadian Community Health Survey

CDC: Centers for Disease Control and Prevention

IPD: Individual patient data

NHIS: National Health Interview Survey

SHeS: Scottish Health Survey

Declarations

Availability of data and material

Unlinked CCHS public use files are available from Statistics Canada and through Ontario Data Documentation, Extraction Service and Infrastructure (www.odesi.ca). Linked CCHS data are available for use by approved researchers at Statistics Canada and at research data centres (www.statcan.gc.ca/eng/rdc/index). Both linked and unlinked NHIS public use data files are available for download from the CDC website (www.cdc.gov/nchs/data-linkage/mortality-public.htm). Public use SHeS data can be obtained from the United Kingdom Data Service website (www.ukdataservice.ac.uk). Access to linked SHeS data can be requested from the

Public Benefit and Privacy Panel for Health and Social Care (www.informationgovernance.scot.nhs.uk).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Canadian Institutes of Health Research, operating grant MOP-142177. The study sponsor had no role in study design; collection, analysis or interpretation of data; manuscript writing; or the decision to submit for publication. Alastair Leyland is funded by the Medical Research Council (MC_UU_12017/13) and the Scottish Government Chief Scientist Office (SPHSU13).

Table 1. Comparison of traditional epidemiology studies and linked population health surveys

	Population Health Surveys	**Traditional Epidemiology Studies**
Purpose	Population health surveillance	Typically etiological questions
Study base	Community-dwelling population	Specific population subgroups
Sampling method	Population-based sampling	Often convenience sampling
Size	Can be very large	Can be very large
Time Frame	Ongoing; often repeated annually	Varies; days to decades
Content	General and broad	Typically purpose-specific
Information Ascertained	Sociodemographic characteristics, health behaviours, health status, health care use	Sociodemographic characteristics, health behaviours, health events, mortality and disease
Physical Measures	Generally only collected in small surveys or from a subsample	Often include collection of anthropomorphic measurements and biological specimens
Study Type	Cross-sectional (usually)	Longitudinal (usually)
Exposure Ascertainment	Typically only at baseline; self-response	At baseline, often with follow-up; electronic data-capture or chart review
Outcome Ascertainment	Data linkage increasingly performed to add mortality and disease outcomes for longitudinal analyses	Typically active ascertainment of study-specific outcomes
International Scope	100+ countries with health surveys; 5+ linked to outcome data	International collaboration occurs but is difficult
Data Access	Public use datasets are increasingly available	Not usually accessible
Documentation	Easily accessible, and detailed documentation available	General methodology available in peer review publications and reports

Source: Authors compilation

Table 2. Comparison of the Canadian, United States and Scottish national health surveys

	Canadian Community Health Survey (CCHS)	National Health Interview Survey (NHIS)	Scottish Health Interview Survey (SHeS)
Country and population size[1]	Canada, 37.6 million	United States of America, 329 million	Scotland, 5.3 million
Primary purpose	Support national, provincial and intraprovincial health surveillance	Support national health surveillance and track progress toward achieving national health objectives	Support national health surveillance, mainly for cardiovascular disease and associated risk factors
Survey design	Cross-sectional household interview survey	Cross-sectional household interview survey	Cross-sectional household interview survey
Administration history	Biennially from 2001 and annually from 2008	Annually from 1957	1995, 1998 and 2003, and then annually from 2008
Content	Health status, heath care utilization and health determinants	Health status, heath care utilization and health determinants	Health status, heath care utilization, health determinants and biological measurements
Target population	Non-institutionalized Canadians 12 years of age and older	Non-institutionalized population of the United States	Persons living in private households in Scotland
Exclusions	Persons living on reserves, full-time members of the Canadian Forces, the institutionalized population and residents of certain remote regions	Persons in long-term care facilities, on active duty with the Armed Forces and in prison, and U.S. nationals living in foreign countries	Persons not living in private households and residents of certain remote islands
Sampling methods	Multistage area-based probability sampling	Multistage area-based probability sampling	Multistage stratified clustered probability sampling
Administration methods	Computer-assisted personal interviews (CAPI) and computer-assisted telephone interviews	Personal household interviews using CAPI	Personal household interviews using CAPI
Survey length	Approximately 45 minutes	Approximately 60 minutes	Approximately 60 minutes
Modules	Core content + province-specific optional modules + rapid response content	Core content + co-sponsored supplementary questions	Core content + rotating biennial modules since 2008 + biological module from a subsample
Sample size	Approximately 130,000 respondents per cycle from 2001 to 2006; approximately 65,000 respondents annually since 2007	Approximately 30,000 adult respondents annually since 1997	Approximately 7,000 adult respondents per survey until 2011; approximately 4,500 adults annually since 2011

[1]Total adult response rate in included survey cycle

Source: Documentation for the Canadian Community Health Survey, United States National Health Interview Survey and the Scottish Health Survey.

Table 3. Creation of comparable smoking, alcohol, physical activity and diet variables from the CCHS, NHIS and SHeS questionnaires

				Comparability		
	CCHS[1] (Canada)	**NHIS (United States)**	**SHeS (Scotland)**	**CCHS NHIS**	**CCHS SHeS**	**NHIS SHeS**
Smoking status				High	High	High
Never smoker	Has never smoked a whole cigarette (SMKDSTY = 6); Former always occasional smoker with <100 cigarette history (SMKDSTY = 5 and SMK_01A = 2)	Nonsmoker with <100 cigarette history (SMKEV = 2 and SMKNOW = 3)	Has never smoked or used to smoke cigarettes occasionally (CIGST1 = 1 or 2)			
Light smoker	Daily smoker with <20 cigarettes/day (SMKDSTY = 1 and SMK_204 <20); Current occasional smoker (SMKDSTY = 2 or 3);	Daily smoker with <20 cigarettes/day (SMKNOW = 1 and CIGSDA1 <20); Occasional smoker (SMKNOW = 2)	Current smoker (daily or occasional) with <20 cigarettes/day (CIGST1 = 4 and CIGDYAL < 20)			
Heavy smoker	Daily smoker with ≥20 cigarettes/day (SMKDSTY = 1 and SMK_204 ≥20)	Daily smoker with ≥20 cigarettes/day (SMKNOW=1 and CIGSDA1 ≥20)	Current smoker with ≥20 cigarettes/day (CIGST1 = 4 and CIGDYAL ≥20)			
Former smoker	Former daily smoker (SMKDSTY = 4); Former always occasional smoker with ≥100 cigarette history (SMKDSTY = 5 and SMK_01A = 1)	Former smoker with ≥100 cigarette history (SMKNOW = 3 and SMKEV = 1)	Former smoker (CIGST1 = 3)			
Drinking Status				High	Moderate	Moderate
Light drinker or non-drinker	Has never drunk (ALCDTYP = 4)	Has never drunk (ALCSTAT = 1)	n/a			
Moderate Drinker	*Males:* Current drinker who consumes 3 or fewer drinks/week on average (ALCDTYP = (1 or 2) and ALCDWKY ≤3)	*Males:* Current drinker who consumes 3 or fewer drinks/week on average (ALCSTAT = (5 6 7 8) and (ALC12MWK x ALCAMT) ≤3)	*Males:* Consumes 3 or fewer units of alcohol/week on average (DRATING ≤3)			
Heavy Drinker	*Females:* Current drinker who consumes 2 or fewer drinks/week on average (ALCDTYP = (1 or 2) and ALCDWKY ≤2)	*Females:* Current drinker who consumes 2 or fewer drinks/week on average (ALCSTAT= (5 6 7 8) and (ALC12MWK x ALCAMT) ≤2)	*Females:* Consumes 2 or fewer units of alcohol/week on average (DRATING ≤2)			
Moderate drinker	*Males:* Current drinker who consumes more than 3 and up to 21 drinks/week on average (ALCDTYP = (1 or 2) and ALCDWKY >3 and ALCDWKY ≤21)	*Males:* Current drinker who consumes more than 3 and up to 21 drinks/week on average (ALCSTAT = (5 6 7 8) and (ALC12MWK x ALCAMT) >3 and ALC12MWK x ALCAMT) ≤21)	*Males:* Consumes more than 3 and up to 21 units of alcohol/week on average (DRATING >3 and DRATING ≤21)			
Binge drinking				High	Moderate	Moderate
Binge drinker	Consumed 5 or more drinks at least once a week in the last year (ALC_3= (5 or 6))	Consumed 5 or more drinks on 52 or more days in the last year (ALC5UPYR≥52)	Consumed 5 or more units of alcohol on the heaviest drinking day in the last week (D7UT08≥5)			
Daily physical activity	Sum(Number of times activity performed in 12 months x Average	Daily METs from vigorous physical activity (6 MET/hour) and	Sum((Number of times activity performed in last 4 weeks x Average	High	Moderate	Moderate

(METs)	duration of activity, in hours x MET value of activity)/365 (PACDEE)	moderate/light physical activity (3 MET/hour) ([6(VIGFREQW x (VIGMIN/60)) + 3(MODFREQW x (MODMIN/60))]/7)	duration of activity, in hours x MET value of activity)/28) *(See MET values below)*			
Fruit and vegetable consumption (excluding juice and potatoes)	Number of times a day consumes fruit and vegetables – Number of times a day consumes juice – Number of times a day consumes potatoes (FVCDTOT - FVCDJUI - FVCDPOT)	Number of times a day consumes fruit (not including juice) + Number of times a day consumes salad + Number of times a day consumes other vegetables (not including potatoes) (FRUIT + SALAD + OVEG)	Portions of all-sized fruit yesterday + Portions of vegetables yesterday (not including potatoes) + Portion of salad eaten yesterday + Portion of vegetables in composites (PORFRT + PORVEG + PORSAL + PORVDISH)	High	Moderate	Moderate

[1]Variable names correspond to CCHS 3.1

Notes: CCHS = Canadian Community Health Survey, NHIS = National Health Interview Survey, SHeS = Scottish Health Survey, MET = metabolic equivalent of task.

Source: Documentation for the Canadian Community Health Survey, United States National Health Interview Survey and the Scottish Health Survey.

Table 4. Linkage to mortality data

	Males			Females		
	CCHS *N* (%[1])	**NHIS** *N* (%[1])	**SHeS** *N* (%[1])	**CCHS** *N* (%[1])	**NHIS** *N* (%[1])	**SHeS** *N* (%[1])
Total	152,888	27,022	12,305	188,083	35,204	15,900
Linked[1]	134,524 (88)	25,342 (94)	10,273 (83)	161,883 (86)	32,890 (93)	13,379 (84)
Deaths among linked[2]	9,675 (7)	2,973 (12)	1,429 (14)	9,552 (6)	3,368 (10)	1,427 (11)
Person-years follow-up[2]	819,453	214,819	71,246	994,431	282,501	93,150

1. Consented to linkage and were successfully linked.
2. From survey administration to follow-up: CCHS and NHIS follow-up to December 31, 2011; 3. SHeS follow-up to December 31, 2014.

Notes: CCHS = Canadian Community Health Survey, NHIS = National Health Interview Survey, SHeS = Scottish Health Survey.

Sources: Canadian Community Health Survey (2003 to 2008) linked to the Canadian Mortality Database (2011); United States National Health Interview Survey (2000, 2005) linked to the National Death Index (2011); Scottish Health Survey (2003, 2008 to 2010) linked mortality (2014).

References

1. Sanmartin C, Decady Y, Trudeau R, et al. Linking the Canadian Community Health Survey and the Canadian Mortality Database: An enhanced data source for the study of mortality. *Health Reports* 2016; 27(12): 10-18.
2. Gray L, David Batty G, Craig P, et al. Cohort profile: the Scottish health surveys cohort: linkage of study participants to routinely collected records for mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. *International Journal of Epidemiology* 2010; 39(2): 345-350. doi:10.1093/ije/dyp155
3. Ingram DD, Lochner KA, Cox CS. Mortality experience of the 1986-2000 National Health Interview Survey Linked Mortality Files participants. *Vital and Health Statistics Series 2* 2008; 147: 1-37.
4. Mindell J, Biddulph JP, Hirani V, et al. Cohort profile: the Health Survey for England. *International Journal of Epidemiology* 2012; 41(6): 1585-1593. doi:10.1093/ije/dyr199
5. Charafeddine R, Berger N, Demarest S, Van Oyen H. Using mortality follow-up of surveys to estimate social inequalities in healthy life years. *Population Health Metrics* 2014; 12(1): 13. doi:10.1186/1478-7954-12-13
6. National Center for Health Statistics. *2015 National Health Interview Survey (NHIS) Public Use Data Release Survey Description.* Hyattsville, Maryland: National Center for Health Statistics, 2016.
7. Manuel DG, Tuna M, Bennett C, et al. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the Cardiovascular Disease Population Risk Tool (CVDPoRT). *Canadian Medical Association Journal* 2018; 190(29): E871-E882. doi:10.1503/cmaj.170914

8. Manuel DG, Perez R, Sanmartin C, et al. Measuring burden of unhealthy behaviours using a multivariable predictive approach: life expectancy lost in Canada attributable to smoking, alcohol, physical inactivity, and diet. *PLoS Medicine* 2016; 13(8): 1-27. doi:10.1371/journal.pmed.1002082

9. Rosella LC, Manuel DG, Burchill C, Stukel TA. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). *Journal of Epidemiology & Community Health* 2011; 65(7): 613-620. doi:10.1136/jech.2009.102244

10. O'Neill M, Kornas K, Rosella L. The future burden of obesity in Canada: a modelling study. *Canadian Journal of Public Health* August 2019. doi:10.17269/s41997-019-00251-y

11. Rosella LC, Lebenbaum M, Li Y, et al. Risk distribution and its influence on the population targets for diabetes prevention. *Preventative Medicine (Baltimore)* 2014; 58(1): 17-21. doi:10.1016/j.ypmed.2013.10.007

12. Hennessy DA, Flanagan WM, Tanuseputro P, et al. The Population Health Model (POHEM): an overview of rationale, methods and applications. *Population Health Metrics* 2015; 13(24): 1-12. doi:10.1186/s12963-015-0057-x

13. Manuel DG, Garner R, Finès P, et al. Alzheimer's and other dementias in Canada, 2011 to 2031: a microsimulation Population Health Modeling (POHEM) study of projected prevalence, health burden, health services, and caregiving use. *Population Health Metrics* 2016; 14(1): 37. doi:10.1186/s12963-016-0107-z

14. Blettner M, Sauerbrei W, Schlehofer B, et al. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology* 1999; 28(1): 1-9.

doi:10.1093/ije/28.1.1

15. Haidich AB. Meta-analysis in medical research. *Hippokratia* 2010; 14(Suppl 1): 29-37.
16. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. *BMC Medical Research Methodology* 2005; 5: 14. doi:10.1186/1471-2288-5-14
17. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; 309(6965): 1351-1355.
18. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *British Medical Journal* 1997; 315(7121): 1533-1537.
19. Coldrtz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary. *American Journal of Epidemiology* 1995; 142(4).
20. Oliveros H. Heterogeneity in meta-analyses: our greatest ally? *Colombian Journal of Anesthesiology* 2015; 43(33): 176-178. doi:10.1016/j.rcae.2015.06.001
21. Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, 2008.
22. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Statistics in Medicine* 1995; 14(19): 2057-2079. doi:10.1002/sim.4780141902
23. Chalmers I. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences* 1993; 703: 156-163; discussion 163-5. doi:10.1111/j.1749-6632.1993.tb26345.x
24. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health*

Professions 2002; 25(1): 76-97. doi:10.1177/0163278702025001006

25. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal* 2010; 340: c221. doi:10.1136/bmj.c221

26. Béland Y. Canadian Community Health Survey - Methodological overview. *Health Reports* 2002; 13(3): 9-14.

27. National Center for Health Statistics. *Data Files Documentation, National Health Interview Survey, 2000 (Machine Readable Data File and Documentation)*. Hyattsville, Maryland: National Center for Health Statistics, 2002.

28. National Center for Health Statistics. *Data File Documentation, National Health Interview Survey, 2005 (Machine Readable Data File and Documentation)*. Hyattsville, Maryland: National Center for Health Statistics, 2006.

29. Bromley C, Sproston K, Shelton N. *The Scottish Health Survey 2003 Summary of Key Findings*. Edinburgh, 2003.

30. Corbett J, Given L, Gray L, et al. *The Scottish Health Survey 2008 Volume I: Main Report*. Edinburgh, 2009.

31. Corbett J, Dobbie F, Doig M, et al. *The Scottish Health Survey 2009 Volume I: Main Report*. Edinburgh, 2010.

32. Bromley C, Corbett J, Day J, et al. *The Scottish Health Survey 2010 Volume I: Main Report.* Edinburgh, 2011.

33. Feeny D, Kaplan MS, Huguet N, McFarland BH. Comparing population health in the United States and Canada. *Population Health Metrics* 2010; 8(1): 8. doi:10.1186/1478-7954-8-8

34. Murray CJ, Ezzati M, Flaxman AD, et al. GBD 2010: design, definitions, and metrics. *The Lancet* 2012; 380(9859): 2063-2066. doi:10.1016/S0140-6736(12)61899-6

35. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 380(9859): 2224-2260. doi:10.1016/S0140-6736(12)61766-8

36. Levin ML, Bertell R. Simple estimation of population attributable risk from case-control studies. *American Journal of Epidemiology*. 1978; 108(1): 78-79.

37. Manuel DG, Yusuf W, Tuna M, et al. cchsflow - An R package for the harmonization of variables across survey cycles. Open Science Framework. 2019. doi:10.17605/OSF.IO/HKUY3

38. Bergeron J, Doiron D, Marcon Y, et al. Fostering population-based cohort data discovery: the Maelstrom Research cataloguing toolkit. Beiki O, ed. *PLoS One*. 2018; 13(7): e0200926. doi:10.1371/journal.pone.0200926

39. Castillo T, Gregory A, Moore S, et al. Enhancing discoverability of public health and epidemiology research data. *Public Health Research Data Forum* 2014.

40. Wong SL, Shields M, Leatherdale S, et al. Assessment of validity of self-reported smoking status. *Health Reports* 2012; 23(1): 47-53.

41. Global Adult Tobacco Survey Collaborative Group. *Tobacco Questions for Surveys: A Subset of Key Questions from the Global Adult Tobacco Survey (GATS), 2nd Edition*. Atlanta, Georgia, 2011.

42. Christensen AI, Ekholm O, Gray L, et al. What is wrong with non-respondents? Alcohol-, drug- and smoking-related mortality and morbidity in a 12-year follow-up study of

respondents and non-respondents in the Danish Health and Morbidity Survey. *Addiction* 2015; 110(9): 1505-1512. doi:10.1111/add.12939

43. Christensen AI, Ekholm O, Glümer C, et al. The Danish National Health Survey 2010. Study design and respondent characteristics. *Scandinavian Journal of Social Medicine* 2012; 40(4): 391-397. doi:10.1177/1403494812451412

44. Harald K, Salomaa V, Jousilahti P, et al. Non-participation and mortality in different socioeconomic groups: the FINRISK population surveys in 1972-92. *Journal of Epidemiology and Community Health* 2007; 61(5): 449-454. doi:10.1136/jech.2006.049908

45. Struijk EA, May AM, Beulens JWJ, et al. Mortality and cancer incidence in the EPIC-NL cohort: impact of the healthy volunteer effect. *European Journal of Public Health* 2015; 25(1): 144-149. doi:10.1093/eurpub/cku045

46. Gorman E, Leyland AH, McCartney G, et al. Assessing the representativeness of population-sampled health surveys through linkage to administrative data on alcohol-related outcomes. *American Journal of Epidemiology* 2014; 180(9): 941-948. doi:10.1093/aje/kwu207

47. Thomas S, Wannell B. Combining cycles of the Canadian Community Health Survey. *Health Reports* 2009; 20(1): 53-58.

48. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutrition* 2002; 5(6B): 1113-1124. doi:10.1079/PHN2002394

Appendix 1—Metabolic equivalent of task (MET) values used to calculate daily physical activity in the Scottish Health Survey

Activity	**METs**
Heavy housework (HRSHWK)	4
Heavy manual labour (HRSMAN)	4
Walking (HRSWLK1)	3
Swimming (SWIMOCC, SWIMTIM)	3
Cycling (CYCLEOCC, CYCLETIM)	4
Working out at a gym / exercise (WEIGHOCC, WEIGHTIM)	3
Aerobics / keep fit / gymnastics / dance for fitness (AEROOCC, AEROTIM)	4
Any other type of dancing (DANCEOCC, DANCETIM)	4
Running/jogging (RUNOCC, RUNTIM)	9.5
Football/rugby (FTBLLOCC, FTBLLTIM)	5
Badminton/tennis (TENNOCC, TENNTIM)	4
Squash (SQUASOCC, SQUASTIM)	4
Exercises (e.g., press ups, sit ups) (EXOCC, EXTIM)	3
Other (1) (ACTAOCC, ACTATIM)	4
Other (2) (ACTBOCC, ACTBTIM)	4
Other (3) (ACTCOCC, ACTCTIM)	4
Other (4) (actdocc, actdtim)	4
Other (5) (DayExc15, ExcTim15)	4